AFRL-IF-RS-TR-2001-245
Final Technical Report
November 2001

CHALLENGE PROBLEM DEVELOPMENT AND EVALUATION MANAGEMENT

Information Extraction & Transport, Incorporated

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. F106

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2001-245 has been reviewed and is approved for publication.

APPROVED: Craig S Anken

CRAIG S. ANKEN
Project Engineer

FOR THE DIRECTOR: Michael Talbert

MICHAEL L. TALBERT, Maj., USAF, Technical Advisor
Information Technology Division
Information Directorate

If your address has changed or if you wish to be removed from the Air Force Research Laboratory Rome Research Site mailing list, or if the addressee is no longer employed by your organization, please notify AFRL/IFTD, 525 Brooks Road, Rome, NY 13441-4505. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	NOVEMBER 2001	Final May 97 - Dec 00

4. TITLE AND SUBTITLE
CHALLENGE PROBLEM DEVELOPMENT AND EVAULATION MANAGEMENT

5. FUNDING NUMBERS
C - F30602-97-C-0147
PE - 62301E
PR - IIST
TA - 00
WU - 06

6. AUTHOR(S)
Robert C. Schrag

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Information Extraction & Transport, Incorporated
1911 North Fort Myer Drive #600
Arlington Virginia 22209

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

Air Force Research Laboratory/IFTD
525 Brooks Road
Rome New York 13441-4505

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2001-245

11. SUPPLEMENTARY NOTES
Air Force Research Laboratory Project Engineer: Craig S. Anken/IFTD/(315) 330-4833

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This final report describes work performed by Information Extraction & Transport (IET), Inc. on Challenge Problem Development and Evaluation Management for the Defense Advanced Research Projects Agency's (DARPA's) High Performance Knowledge Bases (HPKB) program. HPKB had the objective to develop innovative technologies supporting construction (by knowledge engineers) of knowledge bases, ontologies, and associated libraries of problem-solving strategies. IET was responsible for developing a crisis management (CM) challenge problem (CP) to focus and evaluate HPKB technology. The CM CP's application context was the support of intelligence analysts or their automated agents in interpreting international events.

14. SUBJECT TERMS
Knowledge-Based Systems, Crisis Management, Challenge Problem Development, Artificial Intelligence

15. NUMBER OF PAGES
24

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

DARPA's High Performance Knowledge Bases (HPKB) program had the objective to develop innovative technologies supporting construction of knowledge bases, ontologies, and problem-solving strategies. Under this contract, IET developed and administered a crisis management (CM) challenge problem (CP) to focus and evaluate HPKB technology. The CM CP's application context was the support of intelligence analysts or their automated agents in interpreting international events. IET began each evaluation year by defining and refining a CM CP specification including a domain scenario and additional custom source material, sample questions and answers, parameterized questions encoding a combinatorially large space of possible test questions, and scoring procedures. The HPKB evaluations were run as a "friendly" competition, and the results reported here should be treated in this spirit. IET's major contributions to HPKB were in terms of evaluation methodology, challenge problem development, and challenge problem administration. A visitor to IET's HPKB Web site (http://www.iet.com/Projects/HPKB) can find overview briefings, specification materials, evaluation materials, and evaluation results reports at a scale more comprehensive than is represented in this final summary.

Table of Content

List of Figures

List of Tables

Acronym	*Expansion*
AI	Artificial Intelligence
CM	Crisis Management
CP	Challenge Problem
DARPA	Defense Advanced Research Projects Agency
ET	Evaluation Team
HPKB	High Performance Knowledge Bases
IET	Information Extraction and Transport, Inc.
IT	Integration Team
KB	Knowledge Base
PQ	Parameterized Questions
SAIC	Science Applications International Corporation
SME	Subject Matter Expert
SQ	Sample Questions
TFS	Teknowledge Federal Systems
TQ	Test Question
WMD	Weapons of Mass Destruction
Y1	Year 1
Y2	Year 2

1 Introduction

This final report describes work performed by Information Extraction & Transport (IET), Inc. on Challenge Problem Development and Evaluation Management for the Defense Advance Research Projects Agency's (DARPA's) High Performance Knowledge Bases (HPKB) program. HPKB had the objective to develop innovative technologies supporting construction (by knowledge engineers) of knowledge bases, ontologies, and associated libraries of problem-solving strategies. IET was responsible for developing a crisis management (CM) challenge problem (CP) to focus and evaluate HPKB technology.

IET was supported by subcontractor Pacific-Sierra Research (now known as Veridian Systems). Together IET and Veridian Systems—with occasional consulting from Professor Paul Cohen of the University of Massachusetts—constituted the evaluation team (ET) for the CM CP. The CM CP was posed primarily to HPKB's two integration teams (ITs), known by the names of their lead contractors:

- Teknowledge Federal Systems (TFS); and
- Science Applications International Corporation (SAIC).

HPKB spanned three funding years (U.S. Government Fiscal Years 1997–1999), but only two program—and evaluation—years.

- Year 1 (Y1) ran from June, 1997 through July, 1998.[1]
- Year 2 (Y2) ran from July, 1998 through October, 1999.

From the program's outset, IET maintained a Web site disseminating all of its HPKB products (http://www.iet.com/Projects/HPKB). A visitor can find overview briefings, specification materials, evaluation materials, and evaluation results reports.

As a point of departure, Figure 1 depicts the CP development methodology IET created HPKB.

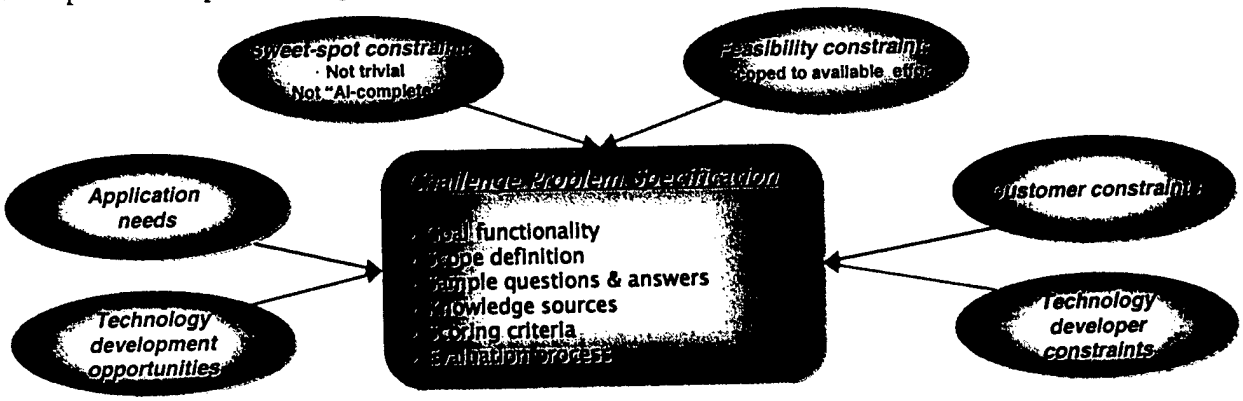

Figure 1: Challenge problem methodology

CPs must merge needs of target applications with opportunities for development of selected technologies into a productive task intersection, tempered by practical customer (DARPA) and technology developer constraints. Hitting just the right level of difficulty requires a thorough understanding of the application domain, the technology, and the reasonably expected pace of

[1] The HPKB Y1 evaluation was featured in the Winter 1998 issue (Volume 19, Number 4) of *AI Magazine:* "The DARPA High-Performance Knowledge Bases Project," by Paul Cohen, Robert Schrag, Eric Jones, Adam Pease, Albert Lin, Barbara Starr, David Gunning, and Murray Burke (pages 25 – 49).

technical development. It is important to set the bar high without making the jump impossible, to make the task feasible without trivializing it.

2 Knowledge representation and reasoning requirements

The CM CP's application context was the support of intelligence analysts or their automated agents in interpreting international events. IET worked with Veridian's subject matter experts to develop an outline of the tasks analysts typically run through when they are tasked by policy-/decision-makers. See the outline below.

I. Information gathering
II. Situation assessment
    A. Explanation
        - Capabilities, motives, intents, risks, rewards
    B. Ramification
        - Effects on actor interests
    C. Context
        - Interests, policies, ideologies, alliances, enmities
III. Scenario development
    A. Action option generation
    B. Option evaluation
    C. Likelihood rating[2]

Information gathering includes tasks corresponding to the journalist's questions "What happened?", "What does it mean?", and "What might happen next?". Situation assessment (or interpretation) includes explanation and ramification factors pertaining to a specific situation at hand and context factors contributing to a "strategic culture" for a national actor's behavior in international relations. Scenario development (or predictive speculation) starts with the generation of plausible actions for each crisis actor. Then options are evaluated with respect to the same factors as in Part II (situation assessment) and a likelihood rating is produced, with the most plausible actions being reported back to the policy makers.

In the tasks of intelligence analysis, there are some classical opportunities for the application of knowledge-based systems. These opportunities (presented as institutional "deficits") are depicted in Figure 2.

[2] This is the only analytic process outline element not posed to be addressed by ITs' development of international political common sense in the CM CP.

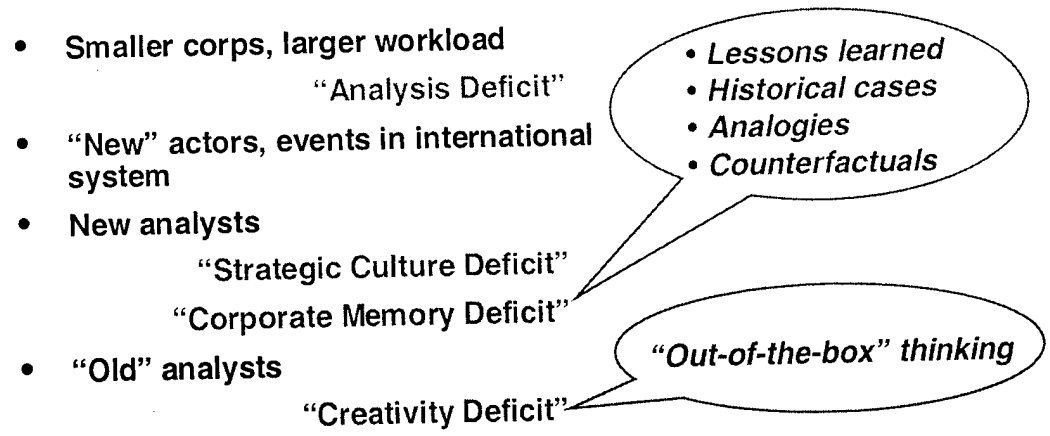

Figure 2: Crisis reasoning objectives

Two main themes in Figure 2 are the use of knowledge bases to retain or multiply corporate expertise and the use of AI-based search to generate analytical possibilities which otherwise might not be considered. The latter usage requires development of extensive common sense knowledge, or "analyst's sense," about the domain to rule out possibilities that are not plausible from an analyst's point of view.As IET reviewed these opportunities with DARPA's Project Genoa leader Admiral John Poindexter, he asked us to concentrate in HPKB Y1 on ways KBs could help analysts break out of their "ruts" of routine analysis ("out-of-the-box" thinking) and in Y2 on how a knowledge-based corporate memory could aid analysts (in ways indicated in the callout bubble).

IET realized that to address these reasoning challenges KBs must capture something akin to "international political common sense," a notion that we depict schematically in Figure 3.

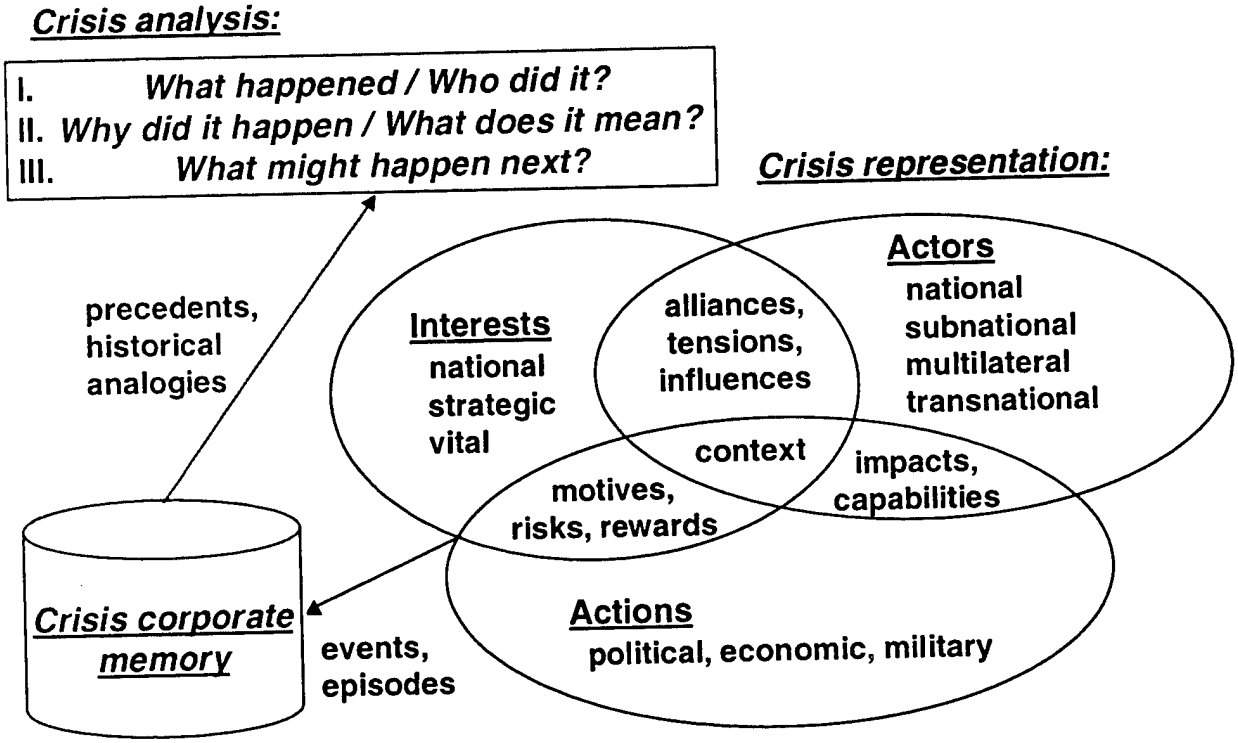

Figure 3: International political common sense

The overlapping ovals in Figure 3 suggest how concepts pertaining to actors, actions, and interests interact. In this model, actions are motivated by interests but balanced by risks and rewards. Actions have impacts and require capabilities. Interests drive the formation of alliances, the exercise of influence, and the generation of tensions among actors. All of these fall against a backdrop of current and historical context.

3 CP specification

IET began each evaluation year by defining and refining a CM CP specification with the following major elements, some of which (those marked "*") we take up in subsections. We refer readers to our HPKB Web pages for a more comprehensive treatment.

- *Domain scenario (crisis storyline) and related historical incident descriptions
- Source material (Web-based and custom-developed background)
- Domain-specific conceptualizations (pre-formal ontologies/KBs)
- Access to subject matter experts (SMEs)
- Sample questions (SQs) and sample answers
- *Parameterized questions (PQs) and supporting PQ grammar, encoding a combinatorially large space of possible test questions (TQs)
- *TQ answer scoring criteria and score aggregation methods

3.1 Crisis scenarios and related historical incidents

IET, in partnership with Veridian Systems, developed Y1 and Y2 fictional crisis scenarios set in the Middle East. Where possible, real events and people were referred to in order to provide both realism and source availability. The exercise of crisis corporate memory in Y2 also required a body of related historical incidents, or "cases" (shown in Figure 5).

3.1.1 Y1 Scenario

The Y1 crisis, which takes place in the Persian Gulf region, involves hostilities between Saudi Arabia and Iran, culminating in closure of the Strait of Hormuz to international shipping. As seen in Figure 4, the Strait of Hormuz forms a strategic chokepoint, less than 40 miles across, through which a large percentage of the world's oil flows. The Iranians currently have missiles that can reach the Strait's shipping channel from Iranian soil—and offshore islands—in less than two minutes. Iran considers its ability to control access to the Strait a political, military, and economic tool. The US, along with Europe and Japan, consider access to the Gulf *via* the Strait of Hormuz a strategic imperative.

Figure 4: Persian Gulf region with Strait of Hormuz

In the Y1 scenario, Iran is vying for hegemonic status in the region, and is critical of Saudi Arabia for its pro-Western stance. Saudi Arabia is extremely wary of Iranian designs on the Gulf and pro-Iranian factions within its borders. The continued inability of the OPEC structure to control oil production exacerbates the situation. The scenario moves through four stages as the conflict escalates. It ends with Iran attacking several Saudi tankers, and declaring the Strait of Hormuz closed to traffic. The result of these actions is a series of armed clashes among several regional powers and United States forces.

3.1.2 Y2 Scenario

In HPKB Y2, the Persian Gulf remained a highly topical setting for a scenario in light of persistent tension and competition among regional actors over economic, security, and sociopolitical interests, the proliferation of weapons of mass destruction (WMD), and ongoing US and Western efforts to bolster stability and ensure the uninterrupted supply of critical energy resources. The Y2 scenario used as its real, current situation ongoing tensions surrounding a dispute over the route of a proposed oil pipeline[3], Iran's economic difficulties, and Iran's well-documented desire to weaken the regional role of Saudi Arabia while enhancing its own. The scenario ended with a fictional but plausible excursion from the real, historical situation.

[3] See map (figure 5) showing the area of interest in Year 2, as well as the proposed pipeline routes.

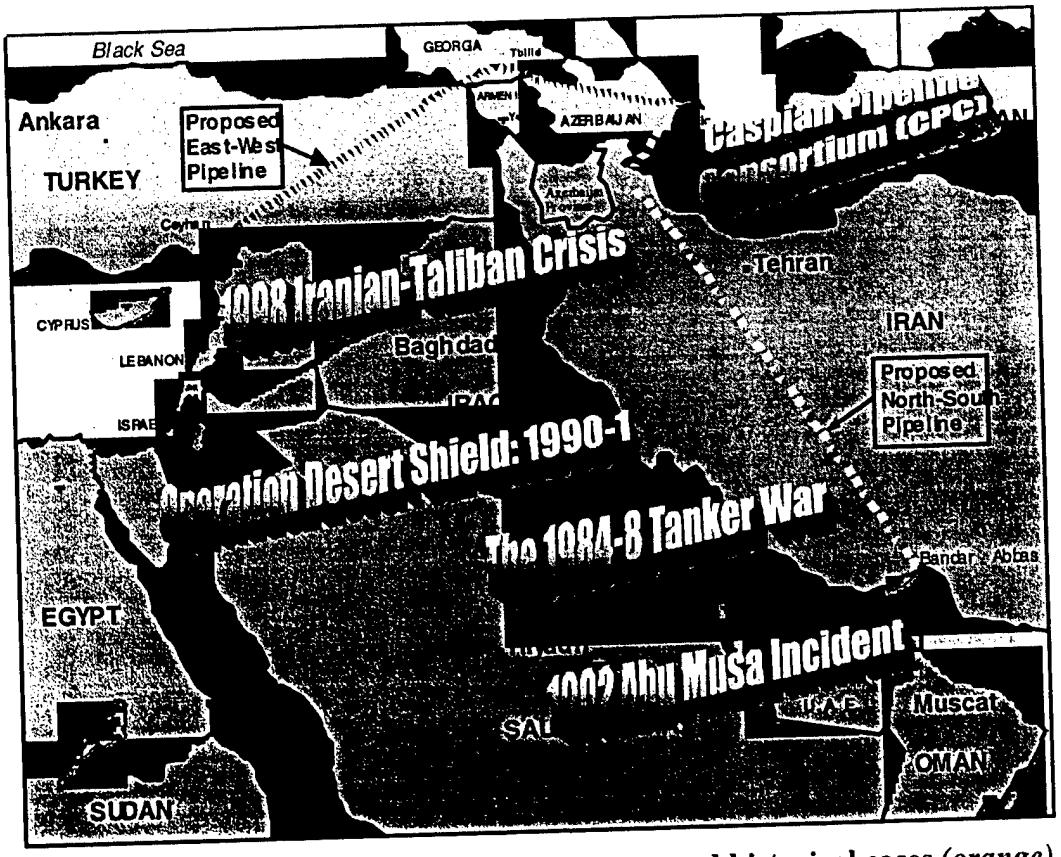

Figure 5: Y2 scenario proposed pipeline routes and historical cases *(orange)*

3.2 PQs

IET developed a KB testing approach based on "parameterized questions" (PQs) that allowed us to generate test questions (TQs) that departed in controlled ways from published sample questions (SQs), giving ITs a well-defined, but largely un-"gamable," target space. Figure 6 demonstrates the PQ notion schematically.

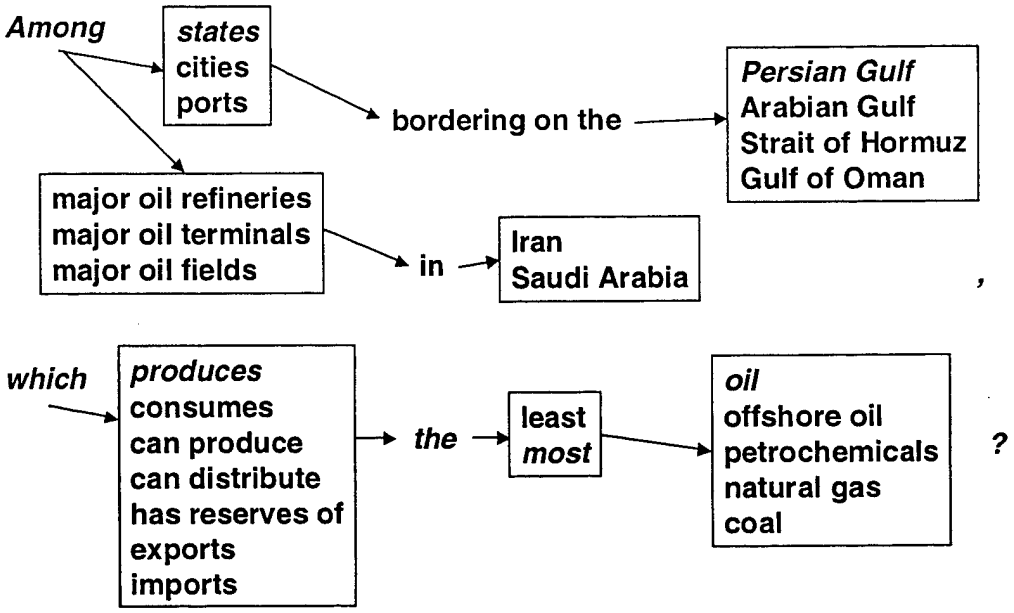

Figure 6. Schematic PQ

By following the arrows and reading the italicized terms in Figure 6, one may find a rendering of CM CP sample question SQ2: *"Among states bordering on the Persian Gulf, which produces the most oil?"*. In each of the boxes are terms which are "ontologically close" to corresponding terms in the sample question. By varying terms in this controlled way, IET is able to produce TQs semantically close to the given SQ, thus limiting practically the scope of knowledge required to be encoded by the ITs, and giving them more indication of what to expect in the actual evaluation period.

Figure 7 presents the Y1 space of "analytical" PQs (those driven by the tasks of intelligence analysis, *vice* simpler PQs intended merely to perform a diagnostic function on the KB—"KB-diagnostic" PQs).

Question type	Sample question	Parameterized template	Number possible
International system • Actors • Actions • Interests	What terrorist group favoring interests of Iran and/or supported by Iran exists within Saudi Arabia?	What <InternationalAgentType> {opposing, favoring} interests of <InternationalAgent2> [and/or supported by <InternationalAgent3>] exists within <InternationalAgent4>?	3.5 billion
Intelligence analysis • Capabilities • Risks • Rewards	What risks would Iran face in exposure of its supporting a terrorist group in Saudi Arabia?	What {risks, rewards} would <InternationalAgent> face/expect in <InternationalActionType>?	1.4 billion
Scenario understanding • Events • Cause • Effect	What is the number of dead caused by the terrorist attack on the oil port during Day 22?	What is the <ScenarioActionResult> caused by the <ScenarioAction> [during <ScenarioTimeInterval>]?	34 million
Background • Economics • Politics • Military • History • Geography	Has Iran ever sponsored a terrorist group performing a terrorist attack?	Has <InternationalAgent1> ever <InternationalActionType> [in <InternationalAgent2>]?	38 million
	What kinds of weapons of mass destruction is Iran believed to possess?	What [kinds of] <MilitaryHardwareType> does/is <Country> {possess, believed to possess, have under development}?	17 thousand

Figure 7: Y1 analytical PQ summary

Figure 7's color coding indicates SQ instantiations of PQ classes and grammar constructs. The combinatorial possibilities for generating syntactically valid TQs were quite large, but in practice the number of acceptable TQs was limited by semantic constraints and by the published source materials.

Our PQ understanding heretofore articulated allows us to provide a more coherent description of the end-to-end KB competencies that we were asking ITs to develop. These competencies are indicated in the sentence that may be read from the *green italic font* in the list below (with CP elements supporting each sentence fragment indicated in the black normal font).

- *Reason in modes of intelligence analysis...*
  - Representative analytical PQs
- *...about limited "situations"...*
  - Crisis scenarios & related historical incidents
- *...based on domain-specific conceptualizations...*
  - International political system
  - Scenario- and case-involved transnational actors
- *...using common sense.*
  - Representative KB-diagnostic PQs

3.3 TQ answer scoring and score aggregation

To accommodate the fact that evaluating TQ answers produced by KBs requires significant subjective judgment, IET devised a discrete 0–3 scoring scale, schematized below. This led to clearer-cut judgments than more wide-ranging scales.

0. **Completely off-target**
1. **Mostly off-target**
2. **Mostly on-target**
3. **Completely on-target**

Figure 8 explains how aggregate scores were computed from raw scores assigned for a given TQ. Individual scores were assigned for criteria falling into four categories: representation, answer, explanation, and source. Each category includes at least one basic and zero or more extra-credit scoring criteria. The basic criteria are used to determine a basic score (left-hand side of Figure 8). Accounting for the extra-credit criteria yields an "overall" score (right-hand side). Here, we account only for the extra-credit criterion "compositionality" (assuming other extra-credit criteria receive 0 scores, *e.g*).

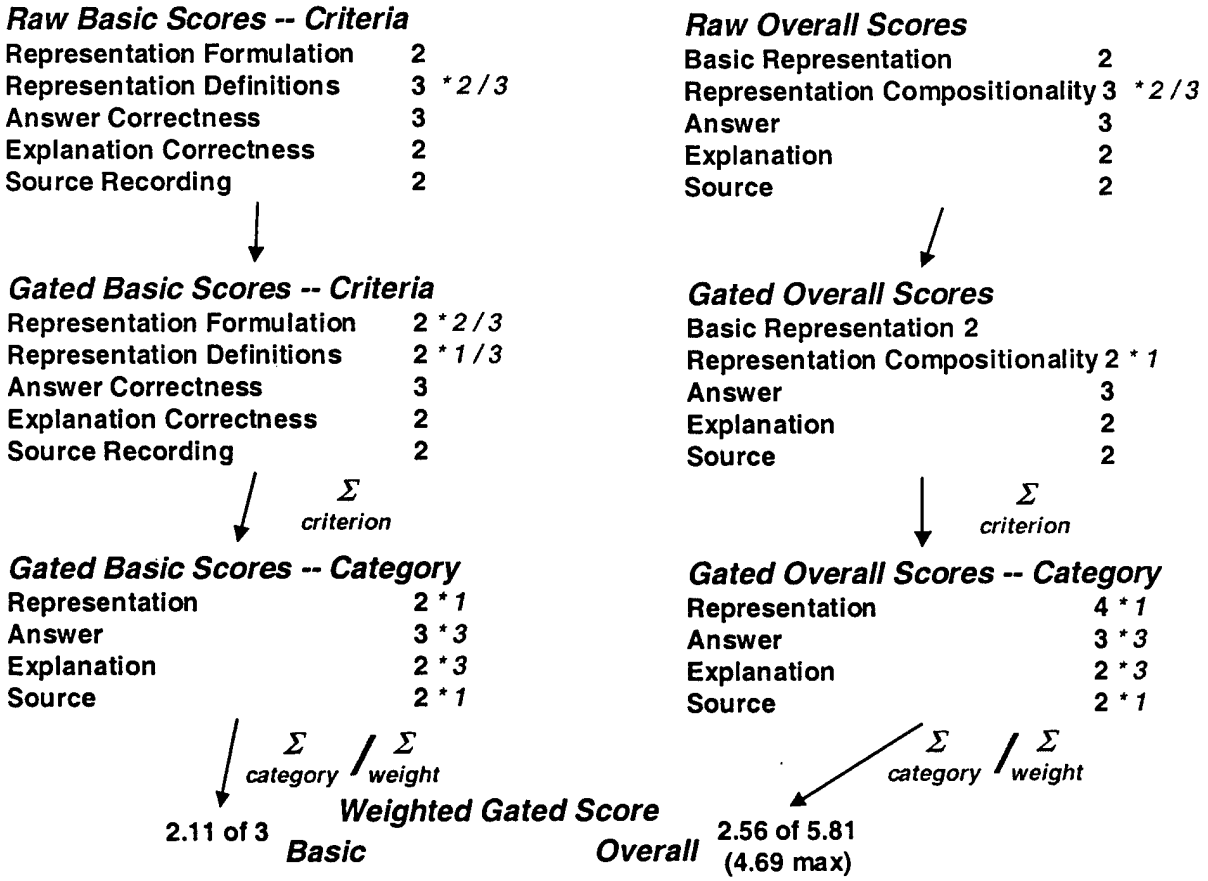

Figure 8: Y2 TQ answer score calculation

Along the left-hand side of Figure 8—basic scores—we have the following, starting from the top.

- Scores for criteria (such as definitions) that are ancillary to a category (here, Representation), are "gated"—that is, reduced by weighting with respect to the category's main criterion's score (here, Formulation) by the fraction of that score over a perfect score (3). This is to prevent the domination of aggregate TQ scores by ancillary criteria (and the possibly accompanying temptation toward gaming). Gating reduces the Definitions score from 3 to 2.

- The same thing occurs for the ancillary extra-credit criterion Compositionality at the top right.

- Next (left, middle), to these gated scores are applied criteria-specific weights to achieve an overall score for the category. (The sum of basic score weights over criteria for a given category always equals 1.) Only the representation category has multiple basic criteria.
  - Compositionality is treated similarly, except that extra-credit criteria weights for a given category do not necessarily sum to 1.
- Next (left, bottom), the category scores are weighted, and the weighted average is used as an overall basic score. (The maximum is 3.)
  - The same weighted averaging is performed for the overall scoring (where the maximum possible and maximum observed scores were 5.81 and 4.69, respectively).

4 Evaluation procedures and results

Figure 9 schematically depicts (not to a natural time scale) the major events involved in the annual evaluation cycle.

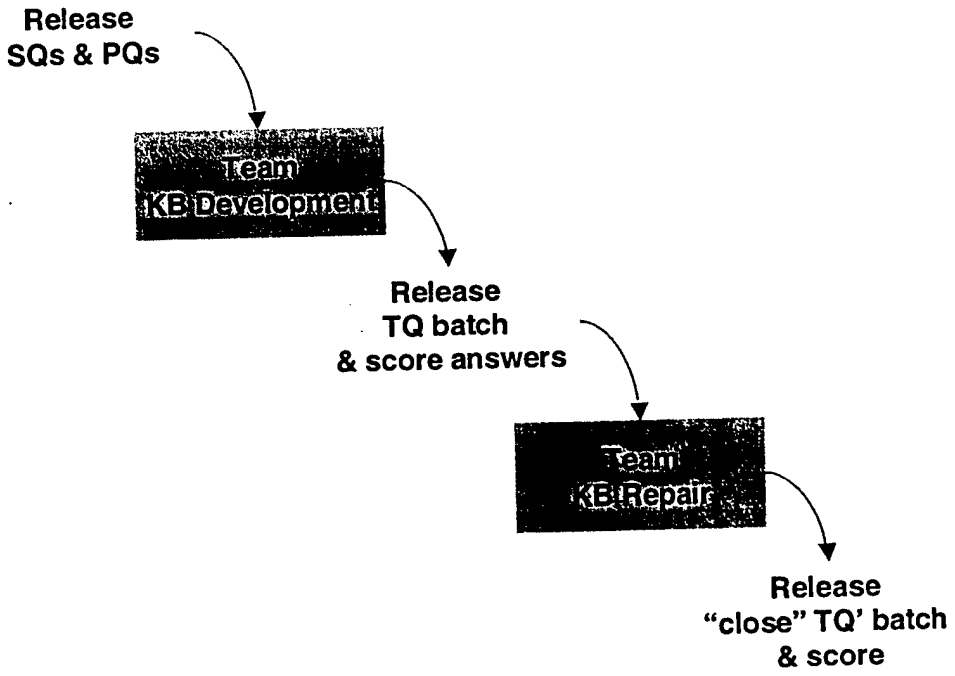

Figure 9: Annual specification-to-evaluation cycle

After the ET's release of a CP spec, ITs would undertake KB and supporting technology development. The start of an evaluation year-culminating, end-to-end evaluation—which lasted only a couple of weeks during the early summer—was marked by the release of "baseline" TQ batches. ITs would have a limited amount of time (less than a day) to elicit their KBs' responses to these batches. The ET would then score ITs' answers, and ITs, using the ET's scores as feedback, would repair (given a day or two) their KBs. This repair was in preparation for response to a subsequent batch of TQs that the ET had generated (on a one-for-one basis) from

the same PQs as had been TQs in the baseline batch. The ET would then score the repaired KBs' TQ answers for the baseline batch (now referred to as a "repair" batch to distinguish the scores) and for the new, robustness-checking, or "close," TQ batch. In all cases, KB testing was "hands-off"—no modifications to the KB were allowed during TQ answering.

Figure 10 shows basic scores, averaged over TQs for the batches administered during HPKB Y1 and Y2.[4] All of the cross-IT scoring differences are statistically significant (based on a paired *t*-test over TQ answer scores) except for the final Y2 batch—TQF.

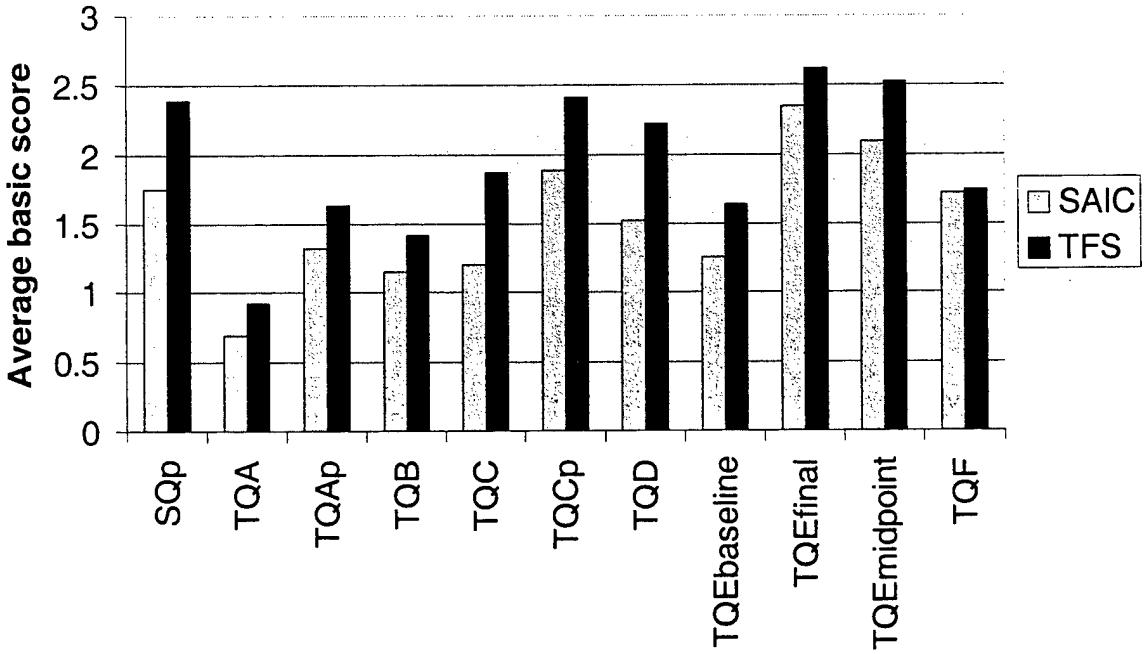

Figure 10: Y2 scores with re-aggregated Y1 scores

The batches noted in Figure 10 were related according to Table 1.

Year, phase	*Baseline batch*	*Repair batch*	*Close batch*
Y1, Phase 1	TQA	TQAp	TQB
Y1, Phase 2	TQC	TQCp	TQD
Y2	TQEbaseline	TQEmidpoint, TQEfinal	TQF

Table 1: Evaluation TQ batch relationships

The HPKB evaluations were run as a "friendly" competition. The graph in Figure 10 (and others like it available on IET's HPKB Web pages) should be treated in this spirit. IET's major contributions to HPKB were in terms of evaluation methodology, challenge problem development, and challenge problem administration.

[4] Y1 basic scores were calculated using a slightly different method. In Figure 10, we have re-aggregated individual Y1 TQs' scores with normalizations to support using the Y2 method.

MISSION
OF
AFRL/INFORMATION DIRECTORATE (IF)

The advancement and application of Information Systems Science and Technology to meet Air Force unique requirements for Information Dominance and its transition to aerospace systems to meet Air Force needs.